Future-Proofing the Web: What We Can Do Today

16 September 2005

John Kunze, California Digital Library

Conclusion

- *Desiccation* now; migration/emulation maybe
- Modest preservation budgets, competing organizational priorities, and diminishing expert format knowledge may make it worth while to save original formats along with simple, low-tech, *desiccated* formats having an excellent preservation outlook just in case
  - the original format should fail and
  - we never get funding to touch the objects again

California Digital Library (CDL)

- A university library with no books, students, or faculty
- Central services for the 10 campus libraries of the University of California
- Content hosting: electronic texts, datasets, finding aids, etc.
- New preservation challenge for CDL: capture and long-term retention of material found on the web

What's digital preservation?

- Stored objects that remain usable and faithful to the creators' original intention
- How? By safeguarding information's …
  - Viability (intact bit streams)
  - Renderability (by machines)
  - Understandability (by humans)
- Viability not in scope here

Migration and Emulation

- Migration problems
  - Unknown costs, human review, format errors
- Emulation problems
  - Unknown costs, human review, software IP
- Both try to keep up with or preserve an object's technical context
- An approach to reduce that context…

The Lesson from Paper

- As a recording and display device
  - Can last for 1000 years
- Why this astonishing performance?
  - No technical intermediation required
- What trick can we borrow
  - The simplest technologies to maintain and understand today are the simplest to carry forward and to recreate in the future

Low-Tech Dependencies

- Semantic technology
  - Loss inevitable due to linguistic shifts
- Substitute light source
  - Fire, the lowest tech invention
- Microfilm
  - Light source plus lens -- 500-year old technology

Desiccated Data

- Remarkable lesson from the longest-lived online digital format
  - Plain text archives of IETF internet RFCs
  - High in value, low in features
- Preservation through “desiccation”
  - No fonts, graphics, colors, diacritics, etc.
  - But essential cultural value retained

Hedging our Bets

- Always save the original format
- In addition, derive desiccated formats in case the original format ever fails
- Extra storage cost may be incurred anyway if your access system requires a plain text derivative for search indexing
- Question: what about Latin-1 support
- Question: surfacing hidden features

Next Lowest-Tech Technology

- Raster image as alternate desiccated format
  - Rectangular grid of picture elements
  - Technical impact of pressure to compress
    - Open run-length encoding or wavelet?
- Rendering tools will never be better than at peak of format's popularity
  - Very common malformed format instances
- Additional fall back format in case the original and plain text versions fail
- Question: surfacing hidden data

Beyond Text and Images

- No attempt yet to formulate desiccated data versions of audio, video, or multi-media
- General lesson: technology will clearly be part of digital preservation, but the greater the technological dependence, the greater the risk

Summary

- Save the original *and* desiccated versions as fall backs in case of failure
  - Few features, much value
  - Low cost, done at peak of tool sophistication
- Web archiving has no preservation metadata
- We may never have the money to touch most of our objects again
- Desiccation is something we can do today

Impersistence - lesser factors

Deliberately or accidentally, objects are

- Removed
- Replaced
- Moved without setting up a redirect
  - Everyone has an indirection mechanism, though most don't use it

- Scheme impact: zero

Impersistence - small factors

Your org likes persistent ids in principle, but

- It lacks knowledge that vanilla web servers trivially support 500,000 redirect directives
- It lacks the expertise or staff to maintain a web server, a two-column database table, and a nightly server config file report writer

- Scheme impact: zero

Scheme costs and risks

- Every modern service needs to support indefinitely and find or be given replacements for at least
  - Web server, web browser, and DNS
- In addition, URN, Handle, and DOI resolution need a global proxy or a plugin *for every* access
- ARK could use a plugin, but doesn't need it
- Handle and DOI also require
  - You to maintain an extra local server
  - The community to maintain a set of global servers
- For the CDL
  - Handle and DOI come with highest risk
  - ARK comes with lowest risk

Persistence - indirect factors

CDL's persistence requirements call for an id scheme (not service) connecting users to

- metadata
- whether and what kind of persistence
- sub-object and variant inferences
- core ids on proxy failure (gracefully)
- Scheme impact: ARK provides these
  - A scheme is not a service (DOI is not CrossRef)
  - When choosing a scheme, we wanted to remain independent of extra external service providers

Our Stuff vs Their Stuff

- Persistence can be split into
  - the Our Stuff Problem
  - the Their Stuff Problem
- It makes no sense for CDL to assign persistent ids to Their Stuff
  - Their Stuff can be hugely important to our users, but we don't control it and cannot vouch for it
  - Where we can afford it, we track them with PURLs
- CDL does assign persistent ids to Our Stuff

Distribution of Id Assignment

- Objects ingested in flows from other libraries per submission agreements
- Each object has an ARK after ingest
  - Either it has it already
  - Or we give it one upon entry
- Campuses can mint their own ARKs or rely on our minting service
- Their own campus ARK namespace is theirs to divide up as they wish

Opaque ids with semantic extensions

- CDL dilemma:
  - opaque ids are needed for names that age and travel well
  - Semantically laden ids are helpful in providing many id services
- Hybrid:
  - opaque ids are used to name abstract preservation objects
  - Semantic and sometimes transient extensions address components inside of objects (the set of components evolves over time anyway)